Advanced Frame Distribution Method Using Padding for Link Aggregation between 10GbE Switches

Soong-Hee Lee, Hyoung-Goo Jeon, *Member, KIMICS*

Abstract—The link e daggregation between 10GbE switches requires an advanced framistribution method to be properly and efficiently applied. The fixed or dynamic frame distribution methods, formerly proposed, cannot fully utilize the aggregated links, where the receiving terminal only attaches to a pre-specified link among multiple physical links. A frame distribution method using padding is proposed for the link aggregation between 10GbE switches to solve this problem. We compared the performance of the proposed method with those of the static and dynamic frame distribution methods. As a result, the proposed method shows a better performance when the offered load is below 0.7 and the average length of the frames is longer than 954 bytes.

Index Terms—frame distribution method, link aggregation.

I. INTRODUCTION

The rapid growth of Internet service chokes the user's need to establish the highspeed networks. This also causes a bottleneck on the MAN (metro politan area network) links that concentrate data traffics between buildings. In addition, more bandwidth and high speed links are required to satisfy the high speed Internet users that pour more traffics into the network. 10GbE (10 giga bit Ethernet) is regarded as a good choice for the high speed links in the metropolitan area.

The 10GbE links, however, are not so cheap to be used to the extent of the end points in the network. A way to efficiently use links is required for the deployment of the 10GbE for the MAN. Accordingly, link aggregation, a way to solve this problem, was standardized as 802.3a in IEEE. This technology organizes a logical link consisting of multiple physical links. As a result, a necessary bandwidth can be attained by just aggregating the existing links without additional new links[4][6][7].

Fig. 1 shows a configuration and the procedure of the link aggregation between networking devices[1][3][5]. As shown in Fig. 1, link aggregation can be applied among switches, between a switch and a server, or between a server and a server. The link aggregation between switches is expected to be most commonly used in the MAN environments.

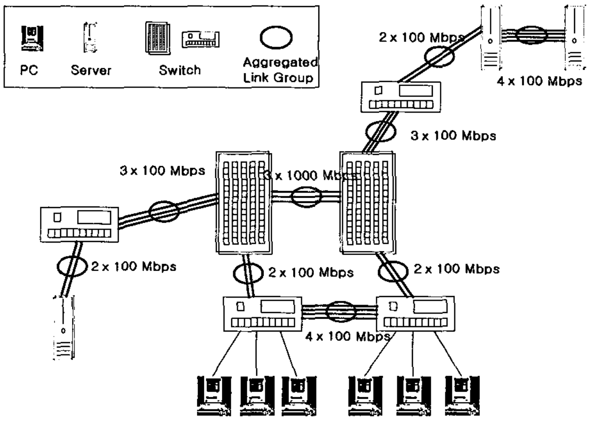

Fig. 1 Example of link aggregation.

Frames are distributed to multiple links participating in the link aggregation. To recover the original sequence of the frames during the link aggregation, an appropriate algorithm for the frame distribution is required before sending frames to the aggregated links. This means that the performance of the devices related to the link aggregation is much dependent on the frame distribution algorithm[6][7]. Therefore, it is important to design a suitable algorithm for the frame distribution among multiple physical links. The detailed algorithm for the frame distribution is not described in the IEEE standards. The standard only describes the followings[1].

- The order of the frames must be maintained during the delivery of the frames into the corresponding ports.
- The duplication of the frames is prohibited.

The frame distribution algorithm for the link aggregation must be designed after considering these two requirements. There have been static and dynamic methods for the frame distribution between switches[2]. These methods, however, cannot fully use the aggregated links at the receiving side. A new way to solve this problem, the frame distribution method with padding between switches, is proposed in this paper.

We presents the problem in the existing frame distribution methods in section II. Section III shows a new method of the frame distribution and compares this with the existing ones. After the computer simulation for the performance comparison in section IV, conclusions are given in section V.

Manuscript received March 30, 2005.

Soong-Hee Lee is with Department of Information communication Engineering, Inje University, Kimhae city Kyoung-Nam Province, Republic of Korea.

Hyoung-Goo Jeon is with Department of Information communication Engineering, Dongeui University, Gaya-Dong, Busanjin-Gu, Busan, Republic of Korea.

II. PROBLEMS IN THE FORMERLY PROPOSED FRAME DISTRIBUTION METHODS

Multiple physical inks are aggregated for the link aggregation, which requires distributing frames from multiple terminals to the multiple links. While there has been no detailed description about the frame distribution in the IEEE standards, a static frame distribution method was proposed[2] which guarantees the ordering of the Ethernet frames after the frame distribution. This method uses MAC addresses of the final receiving terminals as the reference of the frame distribution. Assume M terminals connected to a switch with N links that are aggregated. M/N terminals are grouped to a host group according to the MAC addresses of the receiving terminals and the frames in this group are transmitted through a link toward the terminals. This static frame distribution method is simple and can maintain the order of the transmitted frames through the aggregated links. However, the link efficiency radically declines if frames are concentrated into a specific host since only a link can be used that is fixed by the MAC addresses. In addition, the limited utilization of the links in this method says that this method cannot fully support the true link aggregation. The full utilization of the links is one of the main advantages in the link aggregation.

Another solution, the dynamic frame distribution method, was proposed to solve this problem[2]. This method dynamically reassigns links to disperse the frames which were concentrated into a specific link. A flush buffer is used to prevent the disorder of the frames after the link reassignment. Frames are stored into the flush buffer during the link reassignment and released after the process. Then marginal output links are reassigned to the corresponding host group. The marginal link to be reassigned is selected when the link has the lowest utilization among N-1 output links. This method, however, has complexity in implementation and overhead from repeated reassignments of the links with lower utilization for the increased traffic loads. The overhead may degrade the performance of the switches. Moreover, this method also cannot get the full utilization of the aggregated links at the receiving side, the main advantage of the link aggregation[2].

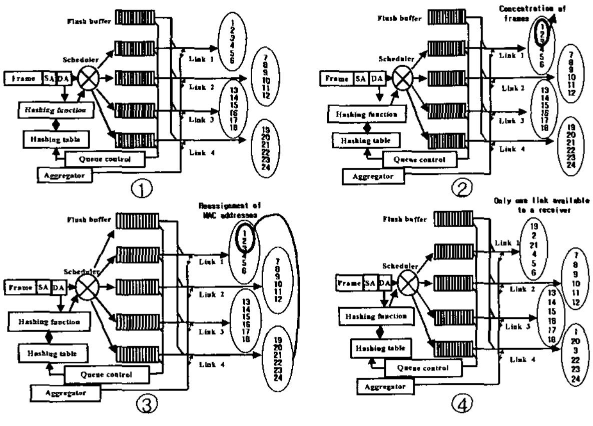

Fig. 2 Problem in dynamic frame distribution.

Figure 2 shows this problem in the dynamic frame distribution method. The problem shown in Fig. 2 is explained as follows:

① Each link or host group is assumed to be mapped to six MAC addresses.
② Link 1 has the heaviest load when frames from terminals with MAC addresses of 1, 2, and 3 are concentrated in that link.
③ Link 4 has the lowest utilization and MAC addresses are reassigned.
④ Only one link among four links can be used in a receiving terminal though the load in the link 1 is lightened after reassigning the MAC addresses. Thus the aggregated links cannot be fully utilized.

III. FRAME DISTRIBUTION METHOD USING PADDING

The formerly proposed static or dynamic frame distribution methods solve the problem of disorder in frame arrivals using the fixed link according to the MAC addresses of the receiving side. Accordingly, these methods cannot get the full utilization of the links applied to the link aggregation, the main advantage of the link utilization. We, therefore, propose a frame distribution method using padding to solve this problem and the disorder problem of frame arrivals.

Frames are sequentially transmitted after padding into the maximum frame length of 1,500 bytes in this method. Thus the disorder of the frame arrivals is solved. Additionally, fixing an output link according to the MAC addresses is not required any more. Hence the method can fully utilize the aggregated links and simpler than the formerly proposed static or dynamic methods. This method originally was applied between terminals which include S/W for the link aggregation, and regarded as unsuitable for switches which mainly relies on H/W implementation[2]. The current use of network processors in the 10GbE switches, however, changes the situation.

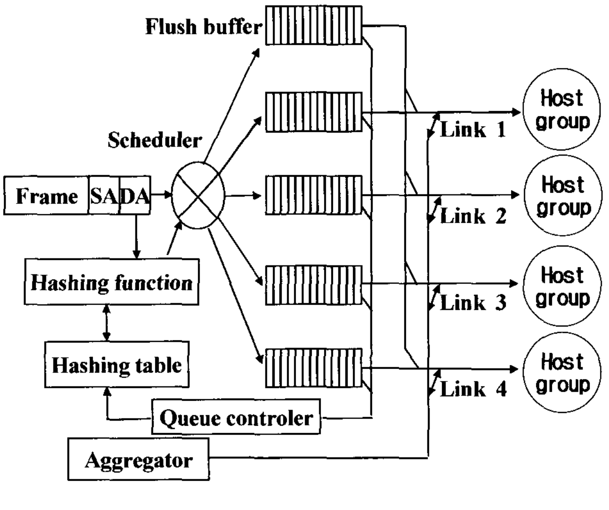

Fig. 3 Block diagram of static/dynamic frame distribution.

Fig. 3 shows an example of the applied static/dynamic frame distribution methods with four links having their own MAC addresses. Hashing function determines the corresponding link for each terminal according to the hashing table. This function, in the dynamic frame distribution method, reassigns the frames, originally assigned to one link, to two links by modifying the hashing table. The queue controller sets a threshold value in the output buffer in the switch. Additional link assignment is started when the buffer level exceeds this threshold value[2].

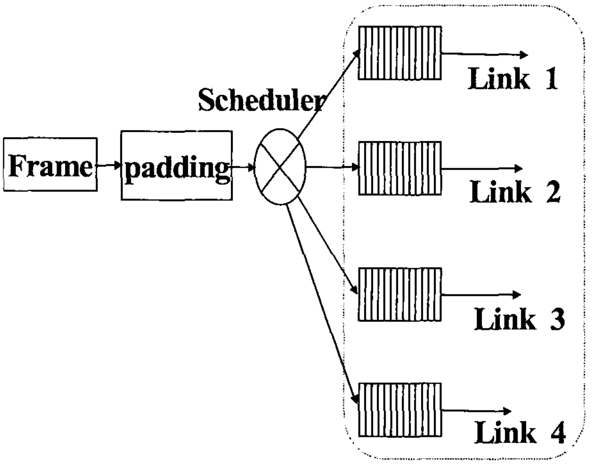

Fig. 4 Block diagram of frame distribution using padding.

Fig. 4 shows an example of the frame distribution method using padding. The method requires no hashing function, hashing table, or queue controller. Frames are padded into the frames with maximum size of 1,500 bytes before sending. Consequently, they are sequentially distributed to the output links to prevent the disorder of the frame sequence. As a result, this method is simpler than the formerly proposed methods. However, the link efficiency is degraded for the smaller sizes of the frames due to the padding overhead.

IV. PERFORMANCE EVALUATION

We compare the performance of the formerly proposed static/dynamic method and the method using padding through the computer simulation.

The model system for the simulation is assumed to have four links between senders and receivers. Each link is assumed to ideally work with no errors during transmission. Lengths of the frames are assumed to accord exponential distribution ranging from 64 to 1,500 bytes. Input traffic is modeled after 2-state MMPP (Markov Modulated Poisson Process) and the total number of frames transmitted during simulation is assumed to be 8,000,000,000. MMPP is applied considering the various types of services in the MAN environments. Frames are assumed to have the maximum frame size after padding and the processed frames are assumed to be sequentially distributed to the four links. The size of the output buffer per link is assumed to be 4 kbytes and 8 kbytes. The receiving side is assumed to have the buffer with the infinite size in each link.

The conditions in the 2-state MMPP used for the simulation are as follows:

- time for staying at state 1 : exponential distribution with the mean of $1/\gamma_1$
- time for staying at state 2 : exponential distribution with the mean of $1/\gamma_2$
- distribution of the customer arrivals at state 1: Poisson process with the average rate of λ_1
- distribution of the customer arrivals at state 2: Poisson process with the average rate of λ_2.

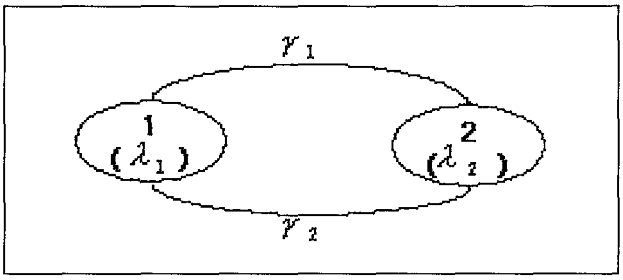

Fig. 5 State diagram of 2-state MMPP

The arrival rate of the frames is attained from the following equation.

$$\lambda = \frac{\lambda_1/\gamma_1 + \lambda_2/\gamma_2}{1/\gamma_1 + 1/\gamma_2} = \frac{\lambda_1\gamma_2 + \lambda_2\gamma_1}{\gamma_1 + \gamma_2} \tag{1}$$

The parameters of the 2-state MMPP used for the input traffic are as follows:

$1/\lambda_1$	$1/\lambda_2$	$1/\gamma_1$	$1/\gamma_2$
5	10	20	25

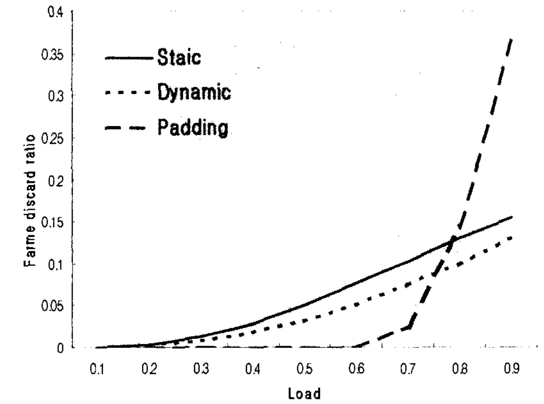

Fig. 6 Frame discard ration of static/dyanmic frame distribution and frame distribution using padding. (buffer size: 4 kbyte, average frame length: 1,086 bytes)

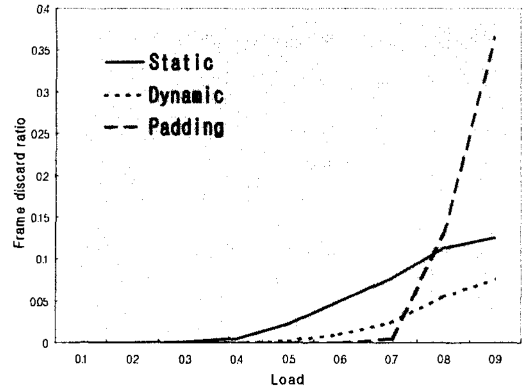

Fig. 7 Frame discard ratio of static/dyanmic frame distribution and frame distribution using padding. (buffer size: 8 kbyte, average frame length: 1,086 bytes)

Fig. 6 and 7 show the frame discard ratio of the static/dynamic frame distribution methods and the frame distribution method using padding. Under the load of 0.7, the method using padding has the lower frame discard ratio with the buffer size of 8 kbyte than the case with the buffer size of 4 kbyte. The method shows similar performance for the other region. The frame discard ratio of the method using padding becomes higher than that of the static or dynamic method under the load heavier than 0.75(for the buffer size of 4 kbyte) or 0.72(for the buffer size of 8 kbyte) even though it has lower discard ratio than the formerly proposed methods under the lighter load. This problem is mainly caused by the increase of the padding overheads of the frames in the output buffers, which tend to increase with the increased load.

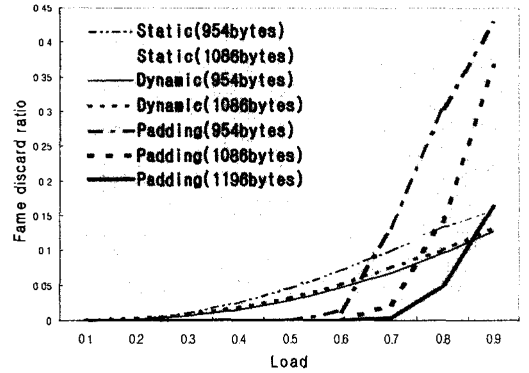

Fig. 8 Frame discard ratio vs. average frame length (buffer size: 4 kbyte)

Fig. 8 shows the frame discard ratio of the static/dynamic methods and the method using padding for the varied average frame sizes of 954, 1,086, 1,196 bytes. The frame discard performance of the static/dynamic method is not sensitive to the changes of the frame size. The method using padding has lower discard ratio under the load lighter than 0.63, for the average frame size of 954 bytes, though it has higher discard ratio under the heavier load. It also has lower discard ratio under the load lighter than 0.75, for the average frame size of 1,086 bytes. For the average frame size of 1,196 bytes, it has lower discard ratio than the static/dynamic methods under the load lighter than 0.85.

Better results with the larger frame size are thought to be caused by the increase of the data amounts which can be processed in the output buffers. Recent trend of the data application using Ethernet is much focused to the transfer of the bulk data like FTP or video services which require more frequent transmissions of the frames with maximum size. This implies that the method using padding can be a good solution for the link aggregation in 10GbE networks.

V. CONCLUSIONS

The existing frame distribution methods and the method using padding are compared in this paper. The former cannot utilize the aggregated links due to the limited use of the output link which is fixed by MAC addresses. The latter can solve this problem by padding method and is simple to implement. Computer simulation results show that the frame distribution method using padding has a better performance with the increase of the average frame size. This means the method can be more efficient to the networks requiring transmission of longer frames such as the file transfer services.

More study is needed for the actual implementation and various situations such as the considerations for link additions and deletions in the aggregated links.

REFERENCES

[1] IEEE, IEEE Standard 802.3, 2000 Edition.

[2] Wu-Jeong Jun, Chong-Ho Yoon, "Performance of Frame Distributions schemes for MAC Controllers with the Link Aggregation Capability", The Journal of the KICS, March 2000.

[3] Link Aggregation according to IEEE 802.3a d White Paper, 2002.

[4] Richard Foote, "Link Aggregation 802.3ad", Corporate Systems Engineering, June 2001.

[5] Walter Thirion, "Link Aggregation Operatio ns", jato Technologies, Inc., July 1998.

[6] IEEE 802.3ad Link Aggregation Task Force Public archive area, http://grouper.ieee.org/ groups/802/3/ ad/public/.

[7] Solving Server Bottlenecks Link Aggregation/FEC/ GEC, http://www.pentium.co.kr/network/connectivity/ solutions/server_bottlenecks/bot_sol2.htm.

Soong-Hee Lee
received the M.S. and Ph.d. degrees in Electronic Engineering from Kyungpook National University in 1990 and 1995, respectively. During 1987-1997. he worked for Electronics and Telecommunications Research Institute (ETRI) of Korea as a member of research staff in the field of ATM networks, B-ISDN and communication network systems. He is now the associate professor of the school of electronics and telecommunication engineering in Inje University. His research interests include communication network systems, high-speed networks, and advanced switching technologies in communication networks.

Hyoung-Goo Jeon
was born in Jeon-Joo Korea on Dec. 1961. He received Bachelor degree at the department of electronics in Inha University in 1987, Master degree and Ph. D. degree from Yonsei University in Seoul Korea in 1992 and 2000, respectively. He is now the associate professor of the school of electronics and telecommunication engineering in Doneui University. His research interests include Digital communication, MIMO-OFDM, WLAN, and IMT-2000.